REFINING SITUATIONAL JUDGMENT TEST METHODS

Peter J. Legree & Joseph Psotka
U.S. Army Research Institute for the Behavioral and Social Sciences
Arlington, VA 22202

ABSTRACT

Situational Judgment Tests (SJTs) assess knowledge, skills, values, and attitudes. They present scenarios, which are based on real events, to be judged, understood, scaled, and interpreted by the examinee. These instruments have been used to evaluate cognitive theories and training programs, and to predict performance. This paper presents the argument that SJTs have potential to renew and reinvigorate many aspects of psychological measurement. We provide a framework to categorize the broad range of procedures and formats adopted for SJTs. The framework indicates that the psychometric range and power of SJTs might be extended by incorporating: (a) Theories and models of human cognition and performance to systematically specify the detail provided in the scenarios; (b) Likert and constructed response formats to maximize breadth of information collected for each scenario; and (c) Consensus-based scoring methods to evaluate knowledge and attitude domains associated with emerging applications.

1. INTRODUCTION

We categorize Situational Judgment Tests (SJTs) broadly as measures designed to assess examinees' opinions and interpretations regarding scenarios that describe or reflect realistic events. These scales have adopted various response formats:

- Frequently using a multiple-choice design and requiring the designation of an action or interpretation as appropriate or inappropriate (Motowidlo, Dunnette & Carter, 1990).
- Sometimes incorporating a Likert scale to appear similar to attitude or survey measures and requiring the assessment of actions or interpretations for each scenario (Wagner & Sternberg, 1985; Legree, Heffner, Psotka, Medsker & Martin, 2003).
- Occasionally providing an open-ended opportunity for examinees to write or voice opinion (Psotka, Streeter, Landauer, Lochbaum & Robinson, 2003).

Table 1 contains an example for readers not familiar with SJTs. For this item, examinees could be requested to identify the most appropriate action, rate the effectiveness of all the actions, or discuss implications of the various actions.

SJTs have been used to construct psychological scales since the 1920's (Moss, 1926). In recent decades, the SJT method has become increasing popular as new conceptualizations regarding work simulation (Motowidlo, Dunnette & Carter, 1990), and tacit knowledge (Horvath & Sternberg, 1986) have broadened goals and range of applications of SJTs. In employment settings, the SJT approach has been evaluated in over 100 studies for personnel selection purposes (i.e., to conduct validity studies and predict performance), and has been shown to have superior validity over traditional techniques using meta-analysis, $\rho = .34$, (McDaniel, Morgeson, Finnegan, Campion, & Braveman, 2001; McDaniel, Hartman & Grubb, 2003). Despite these impressive research findings, SJTs have not been used on a routine basis in the military. *Why not?* Because there are aspects of these measures that break from traditional formats, and an overarching theoretical framework that permits the efficient development of these scales does not yet exist.

Table 1. An SJT example.

A man on a very urgent mission during a battle finds he must cross a stream about 40 feet wide. A blizzard has been blowing and the stream has frozen over. However, because of the snow, he does not know how thick the ice is. He sees two planks about 10 feet long near the point he wishes to cross. He knows where there is a bridge about two miles downstream. Under the circumstances he should:

a. Walk to the bridge and cross it.
b. Run rapidly across the ice.
c. Break a hole in the ice near the edge of the stream to see how deep the stream is.
d. Cross with the aid of the planks, pushing one ahead of the other and walking on them.
e. Creep slowly across the ice.

Unlike conventional knowledge tests, which have been developed to limit uncertainty in item interpretation and have used facts from formal evidentiary sources (Neisser, 1976) with questions such as "*How many miles is it to the moon?*", SJT items have necessarily contained ambiguity because they have simulated complex real-world events and situations that are not yet codified into formal knowledge as rules, dogma, or doctrine. Because SJT scenarios and alternatives are ambiguous, examinees

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **01 NOV 2006**	2. REPORT TYPE **N/A**	3. DATES COVERED **-**

4. TITLE AND SUBTITLE **Refining Situational Judgment Test Methods**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Army Research Institute for the Behavioral and Social Sciences Arlington, VA 22202**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S) 11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES

See also ADM002075.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**8**	

must make inferences in order to respond, and these instruments may appear similar to surveys or even projective tests. However, even when they appear similar, the psychometric properties are quite different. Projective instruments have been developed explicitly to amplify ambiguity based on expectations that examinee inferences and responses will reflect covert, latent and unconscious aspects of personality (cf. Anastasi & Urbina, 1997; Kaplan & Saccuzzo, 1997; Frank, 1939), while SJT applications have been developed to reduce ambiguity and converge on consensus by simulating actual events that have an effective array of responses and can be objectively scored. Surveys, like projective tests, have usually focused on self-report of attitudes and opinions with no known correct answer, but consensus-based methods can establish an objective standard to score even these instruments. SJTs thus represent a blending of assessment methods, incorporating ambiguity, requiring projection and sometimes including a Likert response, yet reflecting both formal and episodic knowledge, which can be scored as a maximal performance scale.

A close inspection of the example in Table 1 illustrates an unusual characteristic of SJT items: all the responses may be correct given reasonable interpretations of the stem and no choice is completely wrong; and at the same time, all the responses may be incorrect given other interpretations and no choice is always right. This characteristic creates novel psychometric problems. While conventional knowledge tests may be deductively scored by using well-regarded theory and knowledge sources, developing scoring standards for SJTs requires consideration of the ambiguities in the stem scenario and the multiple answer options, as well as the complexity of the instruction sets provided to the examinees. Consequently, there are many types of scoring standards for SJTs, and each type has unique psychometric properties. For SJTs that reference formally-described situations, technical or historical documentation might provide the information used to score the answers. However, most SJTs describe complex situations with survey-like options that can only be answered by experts or the consensus of informed groups. The multiplicity of possible standards raises its own problems of deciding which standard is best for which purposes.

In this paper, we attempt to amplify the power of this new technology by clarifying its methods and identifying assumptions and implications that may not be readily apparent with this approach. We conceptualize SJTs broadly as scales requiring respondents to provide opinions about hypothetical situations based on real events, with those opinions compared to expert judgment as opposed to widely accepted fact. We review the strengths and weaknesses of alternate approaches, consider methods used to identify test domains and item content, and explore presentation issues and scoring approaches. Finally, we synthesize our understanding to create a framework to guide the construction and use of SJTs.

2. APPROACH

2.1 Common Components of an SJT

An overview of the four components of SJTs that will be examined throughout this paper follows:

1. Scenario: The dominant innovation of SJTs is the brief description, simulation, image or movie that sets the context for the questions, which we call the scenario. Because the scenario can be very short, it is possible for an SJT to look entirely like a traditional test item, pattern recognition stimulus, or even a survey question. However, the scenario can take much more complicated forms and can vary along many dimensions, primarily level of abstraction, subjective-involvement, and knowledge domain.

2. Response Alternatives: The options attached to each scenario often look like test items that use a traditional, multiple-choice format, although additional data can be collected by incorporating Likert-based or constructed response formats. The new element is that they can deal with broad ranges of intellectual components: knowledge, skills, values, and attitudes. Responses can vary across levels of abstraction, and they can ask about pre-existing conditions, corollary activities, or consequences of the scenario.

3. Judgment Requested: The scenario can be associated with the response alternatives in many different ways, and this is one of the great strengths of SJTs. The instructions can focus on judging how things are, how they should be, or how an individual would actually respond. The judgment can appear similar to a traditional knowledge query concerning whether a response is true; or the judgment may reference values, effectiveness, frequency, likelihood, or any other dimension to elicit useful information. The judgment can deal with subjective or objective alternatives. It can ask about one's self or others. However, probably the most useful and unusual implication is that the choice of judgment may result in the instrument appearing to be a survey and not a test at all (Legree, Martin, and Psotka; 2000).

4. Scoring Standard: The use of extended scenarios with questions that are fundamentally ambiguous and have no single, obvious right or wrong answer has created a powerful innovation in SJTs, the consensus based scoring standard. All assessment instruments provide measures of agreement; but how the standard is set for an instrument is crucial to its validity. Traditional deductive standards can be used, but inductive standards reflecting agreement among either experts or knowledgeable respondents through consensus based scoring algorithms (e.g., difference scores, correlations, or factor analysis) have often proved to be more useful, so far.

2.2 Specifying SJT Content

The development of most SJT scenario and response alternatives has required leveraging a variant of the critical incident technique in which experts report significant events in their area of expertise. These reports are highly objectified stories concerning a situation or activity that are constructed so that the purpose, intent and consequences of an activity must be sufficiently clear to allow an impartial observer to make objective and definite inferences about an activity's outcome or the individuals described involved in the activity (Flanagan, 1954). The critical incident approach has required waves of empirical data collection, and based on these expert data, the inductive identification of relevant scenarios and test items.

We note that SJTs have also been created by using a model of human performance deductively to guide scale creation, and below we argue that this approach can efficiently produce conceptually relevant SJTs. SJTs in this way can be used to evaluate and refine these performance models and simulations. These empirical and model-based approaches parallel inductive and deductive reasoning to the extent the first method has involved collecting descriptions of specific events to identify standards or rules of behavior, while the latter approach has proceeded from the use of more formal frameworks. By analyzing common events, inductive approaches might be able to identify important, recurring situations, while deductive methods might provide better access to rare yet critical relationships.

Inductive Methods. The collection of critical incident data for SJT item production has usually required three phases of data collection with groups of experts or senior job incumbents surveyed to develop representative scenarios, options attached to those scenarios, and scoring rubrics (e.g., Motowidlo et al., 1990; McDaniel and Nyguyen, 2001; McDaniel et al., 2001; Weekley & Ployhart, 2006). In the first phase, groups of experts have been convened for several hours and have been requested to provide critical incidents with instructions that reference general performance or specific competencies. These critical incidents have then been categorized and edited by test developers to create descriptions of representative scenarios. In the second phase, additional experts have been convened in groups for several hours, presented with the phase one representative scenarios and requested to describe and evaluate actions they would conduct if confronted with the described situations. Because these draft items have not reflected doctrine or well-specified knowledge, such as training manuals or position descriptions, usually a third phase of data collection has been required to develop scoring rubrics. This last phase frequently requires surveying experts, although examinee responses have also been analyzed to develop empirical and consensus based scoring standards (McDaniel and Nyguyen, 2001). This process is very time-consuming, easily requiring many months of data collection and analysis to develop a judgment scale.

Apart from cost and time constraints, subtle, yet important limitations with this technique result from requirements that: (i) the actions and consequences described in a critical incident be linked with a high degree of certainty, and (ii) the critical incidents describe only observable events so that an analyst may infer a relationship. These perspectives do not encourage opinion, caveats, or suspicions to be voiced by the observer, and they do not acknowledge the likelihood that actions may have a probabilistic relationship with consequences, and consequences may be determined by multiple antecedents. Moreover, these requirements do not allow the possibility that observer judgments reflect information that is difficult to describe clearly and difficult for an analyst to meaningfully interpret. Instead, this method focuses on easily specified relationships.

We use the term "*probabilistic relationship*" to encapsulate broad, multidimensional relationships, which if fully understood, might provide certainty in assessments of causation. Observations that relationships may be multi-determined and probabilistic can be found in Polanyi (1966), Newell and Simon (1972), and is implied in economic theory developed by Hayek (1948). Moreover and from a philosophical perspective, the empiricist traditions associated with Locke and Hume argued that all inductive reasoning is subject to revision as additional data are collected, and therefore concluded that resultant knowledge is properly considered as probabilistic and not certain. While we agree that complicated relationships may be most fully understood as arising from complex interactions, when those interactions are poorly understood, it may be more useful to accept relationships as probabilistic on a provisional basis and explore the causative substrate as developments allow. To the extent that a judgment of certainty is required from a respondent to link an action to a consequence, probabilistic relationships will be difficult to identify, many of which are complex, important and even critical to our survival. Because the critical incident technique does not identify probabilistic relationships, even on a provisional basis, we believe this method is inherently limited for purposes of developing SJTs corresponding to poorly specified knowledge domains.

While these limitations reflect inherent aspects of the critical incident approach, additional shortcomings with the method are more circumscribed. Within the critical incident workshops, experts have usually been tasked with describing only the most effective, and sometimes the two most effective actions, for each situation (cf. Motowidlo et al. 1990; McDaniel and Nyguyen, 2001). Experts have not been systematically asked to describe ineffective or incorrect responses, which many learning theories have proposed essential to learning and to the assessment of knowledge (e.g., VanLehn, 1990). Instead, ineffective and less-effective response options have often been obtained incidentally to the collection of more-effective options (e.g., Motowidlo & Tippins, 1993), and at worst, these actions have corresponded to less-effective expert responses. Therefore, unlike conventional multiple-choice items on which distracters are factually incorrect, even experts have endorsed options that are used as "distracters" on most SJT items. We know of no data comparing items provided by non-experts against items provided by experts who are simulating non-expertise. However, we note reports that domain experts have been ineffectual in predicting novice performance (Hinds, 1999), and we are skeptical that experts could easily accomplish this task because domain expertise implies rarely committing novice errors: knowing how to do something is much different than knowing the many ways something can be done incorrectly, unless you actually teach novices.

In addition, experts have only been asked to identify responses in reaction to the situation (cf., Motowidlo et al., 1990; McDaniel et al., 2001). Therefore, most SJT items have described problem scenarios with options corresponding to responses that might be followed to rectify the problem. Critical incident writing instructions have not been formulated to identify proactive strategies that might have avoided the situation described in the scenario or to describe interpretations that are more speculative. Such proactive strategies since they require a deeper understanding of the problem situation, might assess expertise more effectively than simply identifying reactive corrective steps.

The approach also runs counter to findings that domain experts are often unaware of the basis of their skill, frequently have difficulty enunciating the basis of their decisions, and the emphasis on actions contrasts with observations that experts expend much effort analyzing a situation before acting, with resultant actions being relatively automatic (Chi, Glaser & Farr, 1988). Moreover, in many domains, expertise has been associated with problem avoidance as opposed to problem reactance, and experts have been expected to perform quickly, solving tasks in real-time, thereby minimizing costs while justifying expenditures. We do not dismiss the critical incident approach, but caution against assuming universal applicability and note the approach runs counter to instructing experts to use all their expertise and experience when providing guidance.

Deductive Methods. A much different approach to constructing tests requiring situational judgment deduces the content of these scales from existing theories and models developed to describe human performance within specific domains. SJTs developed by referencing theories or models have measured emotional intelligence (Mayer Caruso & Salovey, 1999), driver safety (Legree, et al., 2001), social intelligence (Legree, 1995), *psychometric g* (Legree et al., 2000), and temperament (James, 1998). All of these scales match the broad SJT description provided above, although the authors have frequently used other terms to describe these instruments, (e.g., Emotional Intelligence Scales, Tacit and Unobtrusive Knowledge Tests and Conditional Reasoning Scales).

While others have not fully considered the utility of this approach (Weekley & Ployhart, 2006; McDaniel et al., 2001), we believe a summary of models and measures that have been developed using this deductive method suggests many possibilities that might be applied generally to develop SJTs. Because behavioral models often reference antecedents as well as consequences of events, deductively-derived SJTs reflecting these models may sometimes be more inclusive of these factors than those scales that are based on inductive, critical incident based methods.

<u>Psychometric g</u>. Legree, Martin and Psotka (2000) recognized the widely accepted assertion that measures of verbal and general knowledge are highly loaded on *psychometric g* (cf., Carroll, 1993; Jensen, 1980) and gave it a fundamental twist. As is well known, much knowledge loads on *psychometric g*, with very high loadings associated with verbal knowledge. Furthermore, *psychometric g* theory suggests assessment of divergent knowledge can be used to closely approximate *g* (Jensen & Weng, 1994). Many more such models exist in the psychological literature, although few theories may be documented as well as *psychometric g*.

Legree, Martin and Psotka (2000) proposed that *psychometric g* could be measured unobtrusively through survey-like scales requiring judgments. They subsequently created scales that requested individuals to estimate word frequency, identify knowledge implications, and approximate employment distributions.

The items were carefully identified to lack objective referents, and therefore the scales required respondents to provide judgments that were then scored against broadly developed, consensual standards. Performance on this judgment battery correlated approximately .80 with conventional measures of *psychometric g* and the overall results attest to the value of measures that may be readily realized by using a well-documented model of human performance to inform the development of judgment scales. Notice, however, that unlike mathematics or physics questions, the selection of scenarios and options to assess *psychometric g* were guided roughly by theory, but the scoring keys still had to be pragmatically determined through inductive and consensual methods. One limitation with this study is that it was not designed to address the possibility that all judgment is g-loaded, as opposed to being closely tied to specific performance domains.

Crash Risk. Vehicular accident involvement is unusual because meta-analysis has documented only a minor relationship, .10, between crash risk and *psychometric g* (Arthur, Barrett & Alexander, 1991), a finding implying that crash risk judgments would be largely independent of *psychometric g*. According to existing models of crash risk (Näätänen & Summala, 1974; Näätänen & Summala, 1976), drivers reduce crash risk by increasing task effort, modifying speed and minimizing exposure to adverse driving conditions, such as non-mild and inclement weather, road conditions, and distracting events. These models viewed risk management as dynamic, and thereby explicitly recognized that drivers continually modulate their behavior to suit environmental conditions and internal emotional states. This perspective also acknowledged that individuals adjust their driving style, sometimes inadvertently, in response to social pressures and emotional life events in ways that increase crash risk, e.g., faster speeds, shorter headway distances, an increased propensity to commit traffic violations, and more frequent passing. These models have been supported by ample evidence showing that a compensatory process moderates risk in response to ongoing motivations, existing driving conditions and past experience (cf. Evans, 1991; Summala, 1985). Therefore, while the models might have limitations, they had been accepted as useful in the literature.

Based on these models of driving risk, Legree and his colleagues (2001) developed two SJTs that were oriented toward crash risk. One scale required respondents to rate the extent to which a driver should modify his speed based on the presence of specific driving hazards to maintain safety, and the second scale required respondents to assess the extent to which various conditions have been associated with crash involvement. Analyses of these scales identified meaningful factors, with minimal *g*-loadings as expected, and identified dimensions better drivers consider important to reduce risk (Legree et al., 2001). While the results provided additional support for the model, the analyses also extended the model to highlight the importance of the internal state of the driver to crash avoidance.

Emotional Intelligence. Mayer, Salovey, Caruso and Sitarenios (2003) developed a cognitive model of emotional intelligence (EI) that posited four separate facets corresponding to the perception, management, understanding and use of emotional information. These researchers then used their model of EI to develop scales corresponding to the proposed facets. These scales provided a stimulus-scenario and requested individuals to endorse interpretations or identify actions in response to the situation. Analyses of respondent data assessed the construct validity of the EI scales, thereby demonstrating separate factors corresponding to the hypothesized facets. Admittedly, only three factors were demonstrated as opposed to the four hypothesized facets. So while the EI model may have limitations, the predictive validity of the battery has been established against a variety of mental health criteria, and the measure has been accepted as an industrial standard for performance based EI scales (Brackett, Mayer & Warner, 2004; Schultz & Roberts, 2005).

Conditional Reasoning. Conditional Reasoning tests describe situations and then require respondents to assess interpretations associated with the situations (James, 1998). Production of these scales has required identifying *"justification mechanisms"* that have been theoretically associated with specific personality or temperament traits. These mechanisms were then used to deduce SJT item scenarios. For example, aggressive tendencies were hypothesized to reflect attribution biases, and scenario interpretations linked to attribution biases were postulated to indicate aggressive tendencies. These scales have substantial validity (r = .44) against performance-related criteria (James, McIntyre & Glisson, 2004), and reveal a cognitive basis for personality that can be measured through judgment.

Summary. These results demonstrate that deductively developed judgment tests provide powerful tools to investigate domains that vary widely in their loadings on *psychometric g*. Not only have substantial criterion validities been associated with these scales, but resulting analyses have supported theoretical reformations. Judgments on scales corresponding to driver performance had minimal g-loadings, while scales designed to measure to emotional intelligence and conditional reasoning, had moderate loadings. Finally, the data show that carefully-constructed, deductively-derived judgment scales that are aligned with highly g-loaded domains can accurately assess intelligence. Moreover, these scales were produced without the necessity of domain experts by leveraging existing models and theories of human performance. Instead of treating judgment as inextricably linked to general cognitive ability, these results suggest that deductive methods may be used to create SJTs and reinvigorate psychological assessment for many diverse domains.

2.3 SJT Response Formats

Regardless of the means used to identify SJT content, scale construction decisions regarding methods used to describe scenario detail, the type of information respondents provide or endorse, and the approach used to evaluate these responses will influence SJT validity. SJTs have varied in requiring respondents to adopt either a subjective or objective response perspective, and these perspectives may provide access to different types of information. While most SJTs have adopted a power format, it is conceptually possible to develop speeded judgment scales, and the ambiguity inherent in many SJT items favors the selection of a Likert, or constructed response format and not a traditional, multiple-choice format.

2.4 SJT Rubric Development

The process of scoring most knowledge tests assumes that correct answers exist for the items that can be identified using formal knowledge sources. However, SJTs have usually been scored using standards derived from expert opinion because the scenarios reflect actual events that have not been formally described. Analyses conducted for this project (Legree, Psotka, Tremble & Bourne, 2005), demonstrated that expert scoring standards may be closely approximated by analyzing examinee responses using consensus based algorithms. According to the consensual approach, errors are random and ratings data collected from large samples of respondents that contain a range of expertise can be used to approximate the rating means that would be collected from a substantial number of experts, were they available. This demonstration allows journeyman and examinee responses to be used to develop scoring rubrics and evaluate performance. This approach is particularly relevant to scoring SJTs that incorporate Likert response scales, and the approach has been applied to develop scoring standards for domains that have lacked experts (e.g., Legree et al., 2001). This capability decreases costs associated with SJT development by streamlining their development while expanding the domains for which SJTs may be created to include those for which expertise is only emerging.

3. EMERGING APPLICATIONS.

While it is comparatively easy to identify existing judgment tasks associated with psychological theory, many more models and theories exist that might be leveraged to develop judgment scales for a variety of seemingly, intractable domains, thereby providing methods to evaluate and further psychological theory. We proffer expectations regarding the job-analysis method as a potent source for SJTs with military and commercial implications. A difficult problem for many organizations is the identification and production of job knowledge measures against which to validate personnel selection and classification instruments. To address this issue, Industrial/Organizational (IO) psychologists routinely use job analysis techniques to understand personnel requirements. Generally, IO Psychologists survey experts and long-term job incumbents to quantify the importance of job tasks characteristics (e.g., criticality, frequency, trainability, etc.) and the relevance of various knowledge, skills and abilities (KSAs) to job performance (Campbell & Knapp, 2001). This method reflects expectations that the capability to provide sensible task and KSA ratings reflects job-relevant expertise. To insure comprehensiveness of the method, a sampling procedure is often adopted so that different experts will judge different sets of tasks and KSAs. While disagreements among experts are rarely considered, summary data are computed to identify predictor and criterion domains for which to develop measures.

We believe that expectations relating to job analysis procedures provide a general model to construct job-related SJT items. If it is true that expertise is required to provide sensible, informed opinions on job analysis surveys, then the capability to provide sensible opinion should reflect expertise. It follows that a simple method to create domain-specific SJTs is to collect incumbent ratings using a common subset of tasks and KSAs developed for the targeted job and to evaluate these ratings through comparison with appropriate standards to quantify job knowledge. Thereby, converting standard job analysis tasks requiring judgment into job knowledge measures to serve as a "*very low fidelity*" surrogate for job performance. See Table 2.

Table 2. Job-analysis Based Judgment Test

Based on your experience, how ***frequently*** will each of the following tasks be performed ***monthly*** by MPs in a ***combat zone***. Record your rating next to each task on a scale of 1 to 9.

1.___Secure the scene of a traffic accident
2.___Operate a roadblock or check-point
…
30.___Conduct interviews at a crime scene

This approach represents the deductive identification of SJT content with resultant responses assessed as a maximal performance measure via consensus based algorithms. While this concept requires an empirical evaluation before more can be concluded, affirmation of these expectations carries practical implications for the measuring job knowledge by linking job analysis techniques with assessment in one direct step. Disconfirmation would suggest limitations in our understanding of procedures and assumptions supporting the conduct of job analysis. Limited empirical support for this expectation is implied by meta-analysis results showing greater interrater reliability with job analysis ratings data collected from technical experts than incumbents (e.g., $R_{technicalexperts}$=.81 vs. $R_{incumbents}$=.38 using 5 raters and 100 items; Dierdorf & Wilson, 2003).

Because there were no empirical data addressing our expectation, we constructed scales using these KSA methods, embedded the instruments in an ongoing validation study and collected preliminary data. These scales required job incumbents across four military occupations to rate the importance to performance in their specific occupations of (1) a common set of KSAs and (2) occupation-specific job tasks. Preliminary data are encouraging with reliability estimates ranging from .47 with 5 items, to .82 with 30 items and correlations among the scales ranging up to .65. (Refer to Table 3.)

Table 3. Job Analysis SJT reliabilities (R) and correlations.

Topic	**R x KSA 26 items**	**R x Task (# of Items)**	**KSA x Task correlation**
MP	**.82**	**.82 (30)**	**.65 (p<.01)**
Mec v1	**.75**	**.47 (5)**	**.23 (p<.10)**
v2		**.62 (36)**	**.15 (ns)**
Medics	**.58**	**.54 (30)**	**.56 (p<.01)**
Armor	**.78**	**.64 (23)**	**-.17 (ns)**

While many of these values are acceptable by industry standards, it is more remarkable that these judgment scales, which required only several hours to develop and only a few minutes to administer, were never pilot tested. Because the format of these scales was arbitrarily linked to specific variations in the job analysis model, future iterations of the method will likely have better psychometric properties. While we focused on ratings of task importance and frequency, ratings of task criticality, trainability might constitute refinements to the approach.

CONCLUSIONS

All military services have an increasing need to exploit lessons learned and to develop new doctrine in rapidly changing operational environments. To date, extracting and/or assessing Soldiers' experience-based knowledge gained from participating in military operations has been extremely difficult. This research provides a framework for rapidly designing and scoring SJTs that may be well-suited to these emerging requirements. Scales built with these technologies are being integrated into Army's efforts to develop up-to-date performance measures of critical experience-based knowledge for Soldiers.

REFERENCES

Anastasi, A., & Urbina, S. (1997). *Psychological testing (7th ed.)* Upper Saddle River, NJ: Prentice-Hall, Inc.

Arthur, W., Barrett, G. V., & Alexander, Ralph A. (1991). Prediction of vehicular accident involvement: A meta-analysis. *Human Performance, 4,* 89-105.

Campbell, J. P., & Knapp, D. J. (2001). *Exploring the Limits in Personnel Selection and Classification.* Lawrence Erlbaum: Mahwah: NJ.

Carroll, J. D. (1993). *Human Cognitive Abilities.* New York, NY: Cambridge University Press.

Chi, M.T.H., Glaser, R., & Farr, M.J. (1988). *The Nature of Expertise.* Hillsdale, NJ: Erlbaum.

Evans, L. (1991). *Traffic Safety and the Driver.* New York, NY: Van Nostrand Reinhold.

Frank, L. K. (1939). Projective methods for the study of personality. *Journal of Psychology: Interdisciplinary & Applied, 8*, 389-413.

Flanagan, J. C. (1954). The critical incidents technique *Psychological Bulletin, 51*, 327-358

Hayek, F. A. (1948). *Individualism and Economic Order*. Chicago, IL: University of Chicago Press.

Hinds, P. J. (1999). The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance. *Journal of Experimental Psychology: Applied, 5*, 205-221.

James, L. R. (1998). Measurement of personality via conditional reasoning. *Organizational Research Methods, 1*, 131-163.

James, L. R., McIntyre, M. D., & Glisson, C. A. (2004). The Conditional Reasoning Measurement System for aggression: An overview. *Human Performance, 17,* 271-295.

Jensen, A. R. (1980). *Bias in Mental Testing.* New York, NY: Free Press.

Jensen, A. R., & Weng, L. (1994). What is a good g? *Intelligence, 18*, 231-258.

Jones, R.W. et al. (2000). *Critical Incident Protocol - A Public and Private Partnership*. East Lansing, MI: Michigan State University, School of Criminal Justice.

Kaplan, R. M. & Saccuzzo, D. P. (1997). *Psychological testing: Principles, applications, and issues (4th ed.).* ; Belmont, CA: Brooks/Cole Publishing.

Legree, P. J. (1995). Evidence for an oblique social intelligence factor established with a Likert-based testing procedure. *Intelligence 21,* 247-266.

Legree, P. J., Heffner, T. S., Psotka, J., Medsker, G. J. & Martin, D. E. (2003). Traffic crash involvement: Experiential driving knowledge and stressful contextual antecedents. *Journal of Applied Psychology, 88, 15-26.*

Legree, P. J., Martin, D. E &. Psotka, J., (2000). Measuring cognitive aptitude using unobtrusive knowledge tests: A new survey technology. *Intelligence 28*, 291-308.

Legree, P. J., Psotka J., Tremble, T. R. & Bourne, D. (2005). Using Consensus Based Measurement to Assess Emotional Intelligence. In R. Schulze & R. Roberts (Eds.), *International Handbook of Emotional Intelligence.*

Mayer, J. D., Salovey, P., Caruso D. R., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. *Emotion, 3,* 97-105.

Mayer, J. D., Caruso D. R. & Salovey, P. (1999). Emotional intelligence meets traditional standards for an intelligence. *Intelligence*, 27, 267-298.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braveman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730-740.

McDaniel, M. A. & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9,* 103-113.

Moss, F. A. (1926). Do you know how to get along with people? *Scientific American, 135*, 26-27.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*, 640-647.

Motowidlo, S. J., Tippins, N. (1993). Further studies of the low-fidelity simulation in the form of a situational inventory. *Journal of Occupational and Organizational Psychology, 66*, 337-344.

Näätänen, R. & Summala, H. (1974). A model for the role of motivational factors in drivers' decision-making. *Accident, Analysis & Prevention, 6,* 243-261.

Näätänen, R. & Summala, H. (1976). *Road User Behavior and Traffic Accidents*. Amsterdam: North-Holland.

Newell, A, & Simon H. A. (1972). *Human Problem Solving.* Englewood Cliffs, NJ: Prentice-Hall Inc.

Polanyi, M. (1966). *The Tacit Dimension*. New York: Doubleday.

Psotka, J., Streeter, L. A., Landauer, T. K., Lochbaum, K. E., & Robinson, K. (2004). Augmenting Electronic Environments for Leadership. In Advanced Technologies for Military Training: Proceedings No. RTO-MP-HFM-101-21 of the Human Factors in Medicine Panel, Genoa, Italy, October 13, 2003. Research and Technology Organization, Neuilly-sur-Seine: France. pp. 287-301.

Schulze R. & Roberts R. (2005). *International Handbook of Emotional Intelligence.*

Summala, H. (1985). Modeling driver behavior: A pessimistic Prediction? In L. Evans & R. C. Schwing (Eds.), Human Behavior and Traffic Safety (pp. 43-65). New York: Plenum.

VanLehn, K. (1990). *Mind bugs: The origins of procedural misconceptions.* Cambridge, MA, US: The MIT Press.

Wagner, R. K. & Sternberg, R. J. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality & Social Psychology, 49,* 436-458.

Weekley, J. A., & Ployhart, R. E. (2006). Situational Judgment Tests: Theory, Measurement, and Application. Lawrence Earlbaum Associates: Mahwah, NJ.